OntoGene (Team 65): preliminary analysis of participation in BioCreative III

Rinaldi, Fabio ; Schneider, Gerold ; Clematide, Simon ; Jegen, Silvan ; et al

Abstract: Background: The BioCreative series of competitive evaluations of text mining systems provide a major test bed for novel techniques in biomedical text mining. Results from the previous and current competition are of fundamental importance for further development in the area. Results: The OntoGene group participated in all tasks of the current edition. Preliminary results seem satisfactory, however a detailed analysis cannot be performed without a comparison with the results of the other participants.

OntoGene (Team 65): preliminary analysis of participation in BioCreative III

Fabio Rinaldi*[1], Gerold Schneider[1], Simon Clematide[1], Silvan Jegen[1], Pierre Parisot[2] , Martin Romacker[2] and Therese Vachon[2]

[1] Institute of Computational Linguistics, University of Zurich, Switzerland
[2] NITAS/TMS, Text Mining Services, Novartis Pharma AG, Basel, Switzerland

Email: F. Rinaldi*- rinaldi@ifi.uzh.ch; G. Schneider - gschneid@ifi.uzh.ch; S. Clematide - simon.clematide@cl.uzh.ch; S. Jegen - silvan.jegen@novartis.com; P. Parisot - pierre.parisot@novartis.com; M. Romacker - martin.romacker@novartis.com; T. Vachon - therese.vachon@novartis.com;

*Corresponding author

Abstract

Background: The BioCreative series of competitive evaluations of text mining systems provide a major test bed for novel techniques in biomedical text mining. Results from the previous and current competition are of fundamental importance for further development in the area.

Results: The OntoGene group participated in all tasks of the current edition. Preliminary results seem satisfactory, however a detailed analysis cannot be performed without a comparison with the results of the other participants.

Background

OntoGene is a research project based at the Institute for Computational Linguistics of the University of Zurich, focusing on the usage of advanced natural language processing techniques for the purpose of biomedical text mining. Since the beginning of our activities in this domain (2005), our core focus has been on relation extraction [1], rather than on entity extraction.

We participated in the previous two editions of the BioCreative shared evaluation. In BioCreative II (2006) we had the best reported results in the extraction of experimental methods task (PPI-IMT) and very competitive results in the extraction of protein interactions (PPI-IPT) [2]. In BioCreative II.5 (2009) we obtained the best results (according to the 'raw' AUC metric) in the main task of the competition (extraction of protein interactions) [3, 4].

Due to very recently obtained additional research funding, we decided to increase our effort in the current competition, and participate in all of the tasks on offer. In the rest of this research report we describe in detail our approach to each of the tasks.

RUN 1	RUN 2
Positives: 15101	Positives: 17973
Relevant: 1670	Relevant: 1670
TP: 451	TP: 467
FN: 1219	FN: 1203
FP: 14650	FP: 17506
Recall: 0.2701	Recall: 0.2796
Precision: 0.0299	Precision: 0.0260
Averaged-TAP-5: 0.0718	Averaged-TAP-5: 0.0891
Averaged-TAP-10: 0.0992	Averaged-TAP-10: 0.1073
Averaged-TAP-20: 0.1077	Averaged-TAP-20: 0.1156

Table 1: GN results on the 50-articles evaluation set

Results and Discussion

GN Task

In the GN task we used a variant of the OntoGene text mining system which was previously developed for the detection of protein-protein interactions. While the full OntoGene system includes modules for syntactic parsing and relation extraction, the version used for the GN task included only part of the complete pipeline. The following processing steps are performed: (1) XML cleanup and transformation into our own basic XML format; (2) preprocessing with Lingpipe [5] (sentence splitting, tokenization, tagging); (3) terminology recognition; (4) detection of 'focus organisms'; (5) terminology filtering and scoring.

The terminology recognition module is based on an efficient lexical lookup approach, with the contribution of a 'normalization' module (rule based) which can take into account the most frequent surface variants of a term. The lookup uses an internal terminological resource built using terms extracted from UniProt, Entrez Gene, NCBI Taxonomy, Cell Line Knowledge Base (CLKB). An additional aim of our participation was to test an extensive gene resource provided by TMS (Text Mining Services, Novartis AG, Basel).

One characteristic of our approach is the usage of a specific module for the detection of the 'focus organism', i.e. the core specie(s) discussed in the paper. This information is later used for the disambiguation of gene and protein mentions. This module was originally optimized for disambiguation of protein mentions over the set of IntAct 'snippets'.[1] No further adaptation for the GN task in BC III was performed

We use a terminology filtering and scoring approach, which is based on the one hand on textual features, on the other hand on the detected organism. It functions as follows: for each term for which a focus organism above a probability threshold filter has been identified, and which is not in a stop word list, a score based on frequency of the term, the zone (title, abstract, main text), and organism-related keywords is calculated. Organism-related keywords express e.g. that the presence of the word 'murine' gives increased scores to terms related to mouse. The scores and the organism-related keywords were manually adapted to the training documents. Broadly speaking, for each term candidate $SCORE = f * org$, where

f : frequency of term in text (an occurrence in the title has a weight of 200, an occurrence in the abstract a weight of 8; additionally terms in italics are weighted 3 times higher).

org : organism score from "focus organism" detection module (rebalanced through some specific additional organism-related keywords).

The difference between our two submitted runs is mainly in the terminological resources. RUN 1 does not use EntrezGene or UniProt, but instead used an extensive terminological resources provided by TMS (Text Mining Services, Novartis AG), which however covers only the five most important species (human, mouse, rat, yeast and drosophila). Additionally, we included organism resources extracted from the NCBI taxonomy and terms from the CLKB. The TMS resource contains 670,000 term senses. Our own organism and CLKB resource contains 49,000 term senses. This resulted in 520,000 normalized terms, and 172,000 different gene IDs from 5 different organisms.

RUN 2 additionally used 2,203,000 terms from UniProt (version from June 2010) and 1,021,000 terms from EntrezGene (only 20 topmost organisms from the training data, for efficiency reasons). This resulted in 1,856,000 normalized terms and 833,000 different gene IDs from 2,113 different organisms.

[1] A snippet is a short textual reference provided by the IntAct curators.

ACT	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
TP	351	539	756	648	475
FP	120	353	1823	1317	285
FN	559	371	154	262	435
TN	4970	4737	3267	3773	4805
sensv.	0.38571	0.59231	**0.83077**	0.71209	0.52198
specf.	**0.97642**	0.93065	0.64185	0.74126	0.94401
accur.	**0.88683**	0.87933	0.67050	0.73683	0.88000
Matthew	0.48297	**0.52727**	0.34244	0.34650	0.50255
P at full R	**0.16189**	**0.16189**	0.15182	0.15182	0.15660
AUC iP/R	0.63847	**0.63890**	0.41741	0.41740	0.62394

Table 2: PPI-ACT Performance: specf (specificity), sensv (sensitivity), accur (accuracy).

The results obtained on the 50-articles set released by the organizers after the end of the competition are shown in table 1. Not having seen the results by other teams, the only conclusion which we can draw at present is that the resource used for RUN 1 appears to be sufficiently complete, in comparison with the subset of EntrezGene used for RUN 2. In fact, in RUN 2 we have an increase of only 16 TP (+3.5%), which is small compared with the increase of 2,856 FP (+19.5%). Unexpectedly, the TAP-k measures are definitely better for RUN 2. This would suggest that RUN 2 produced a better ranking than RUN 1. A possible explanation for this difference is that the contribution of the "focus organism" detection module is better in RUN 2 than in RUN 1 (therefore genes belonging to the selected organisms are ranked higher). Our "focus organism" module [6] was initially developed for PPI detection. In order to derive an organism ranking it uses all relevant terminology in the article: in particular terms from NCBI and CLKB, but also proteins mentions. Crucially however, it does not use gene mentions to the same extent as protein mentions (in retrospect, we should have adapted it to the nature of the competition). Therefore the lack of sufficient protein mentions in RUN 1 produced a lower quality ranking of organism, which in turn resulted in a worse ranking for genes.

On the set of the 50 most difficult articles, we reached an unweighted average TAP-20 of 0.07 for RUN 1. On the training data we had reached an unweighted TAP-20 of 0.3453. The low results for the 50 articles set is mostly due to the fact that only 103 gene IDs out of 1,219 false negatives were available in this resource. For RUN 2, we had 1,203 false negatives. However also here, only 335 gene IDs were available in our resource. On the training data we had reached an unweighted TAP-20 of 0.3751.

PPI-ACT Task

Three of the runs were generated applying Maximum Entropy optimization (specifically the software package 'MEGAM' [7]). Features considered include lexical items in the document (+Bow),[2] MeSH annotations (+Mesh),[3] and a score delivered by our PPI detection pipeline (+PPIscore)[4] . Two runs (RUN 3 and RUN 4) used only the result of the PPI pipeline. The development set proved to be representative for the testset.

The feature weights used for the test set were drawn from the development set only. Including the balanced (but therefore biased) training set (which was released earlier in the shared task) proved to detoriate the results in a 10-fold cross-validation experiment on the development set. Using the bow and mesh features, we get a huge number of features. In order to keep the training efficient, and to prevent over-training, each

[2]All words of the articles were stemmed. Than all counts of a stem were used as a feature. E.g, if the word "protein" was found 3 times, we produced the features "protein_1", "protein_2", "protein_3". This produced for instance 70886 different features for the development set.

[3]Every MeSH descriptor, with and also without every qualifier, was used as a feature. E.g., for the MeSH term "-Signal Transduction (-drug effects; +physiology)" as it appeared in the textual format, we produced the descriptor features "signal/transduction/drug/effects", "signal/transduction/physiology". For multi word terms, we added also all descriptor terms produced by iteratively removing the first word, for instance "transduction". Additionally, all MeSH qualifiers as "-drug/effects" and "+physiology" were added.

[4]This feature is computed using the full pipeline for detection of PPI as used in the BioCreative II.5 challenge. The original system is used to detect candidate interactions, and deliver each of them, together with a numerical score. This value was discretized in order to form few large classes and then used as a feature set.

IMT	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
Evaluated Results	5098	21529	4576	666	21600
TP	447	527	431	223	527
FP	4651	21002	4145	443	21073
FN	80	0	96	304	0
Micro P	0.08768	0.02448	0.09419	**0.33483**	0.02440
Micro R	0.84820	**1.00000**	0.81784	0.42315	**1.00000**
Micro F	0.15893	0.04779	0.16892	**0.37385**	0.04763
Micro AUC iP/R	0.27588	0.24484	0.27727	0.14169	**0.29016**
Macro P	0.09346	0.02448	0.09992	**0.33483**	0.02440
Macro R	0.83206	**1.00000**	0.79377	0.42883	**1.00000**
Macro F	0.16322	0.04750	0.17163	**0.35403**	0.04735
Macro AUC iP/R	0.47884	0.44034	0.47650	0.30927	**0.50111**

Table 3: PPI-IMT Performance

feature had to appear at least 3 times in the development set, and additionally, the feature selection limitation of MEGAM was used to allow not more than 20,000 features. The resulting features are distributed as follows: 69% bow, 31% mesh.

RUN 1, as expected was the run with the highest accuracy (see table 2). Specificity was deliberately maximized at the cost of sensitivity because of the class imbalance. The features used were +PPIscore, +Mesh, +Bow with standard class binarization of MEGAM at 0.5 between classes 0 and 1. RUN 2 was aimed at maximizing Matthew's correlation coefficient. It is also the run with the highest AUC. The features used were +PPIscore, +Mesh, +Bow with lowered binarization threshold of MEGAM at 0.2 between classes 0 and 1, in order to boost the positive class (the threshold was determined heuristically on the basis of the development set). RUN 3 was aimed at maximizing recall (without using maximum entropy optimization). The 'raw PPIscore' was discretized as follows: if $PPIscore > 0.2$ then class=1 else class=0. RUN 4 was aimed at a balanced specificity / sensitivity result. It did not use the maximum entropy approach, but only the raw PPIscore with the following decision rule: if $PPIscore > 1.1$ then class=1 else class=0. RUN 5 used only the +Bow and +Mesh features, with lowered binarization of MEGAM at 0.25 between classes 0 and 1, in order to obtain the best Matthew's coefficient (threshold determined by experimentation on the development set). The comparison with RUN 2 is particularly interesting because it shows the impact of the +PPIscore feature: we gain 64 TP, but also get 68 more FP.

We have made the following observations. First, the class imbalance negatively affects the recall of the smaller class (1), because the classifier optimizes for overall accuracy. One way to improve the high recall results might be to use the several subscores that make up PPIscore (for example syntactic path, word at the top of the path, protein pair salience, zoning information, etc.) as fine-grained individual features, whose weights can also be optimized individually.

PPI-IMT Task

For the PPI-IMT detection task, we have developed two statistical systems (called system A and system B in this document). Both are based on a naive Bayes approach but use different optimizations and heuristics. The submitted runs correspond to the following:

RUN 1: full output of system A
RUN 2: full output of system B
RUN 3: optimized output of system A
RUN 4: optimized output of system B
RUN 5: combined output (average scores of RUN 1 and RUN 2)

The full outputs were aimed at maximizing R and AUC, the optimized outputs at maximizing F-score. We have avoided sending runs which optimize precision, because these can always be obtained by picking for each article only the best prediction (i.e. the method which is ranked first). [8] reports that the curators preferred a high recall setting to a high precision setting, because it is much easier and less time-consuming to reject suggestions (false positives, low precision) than to add new information from scratch (false negatives,

$p(method\|word)$		
Probability	Word	Method
0.490056	L1	MI:0006
0.470270	LT	MI:0019
0.447269	ERK1/2	MI:0006
0.443877	hydrogen-bonding	MI:0114
0.441441	omit	MI:0114
0.438765	synapses	MI:0006
0.436363	tumours	MI:0006
0.435114	REFMAC	MI:0114

Table 4: Statistical association of methods with specific words (examples)

low recall). A good ranking, coupled with good recall, allows the user to decide where to stop examining the results, rather than leaving the decision to the system.

RUN 5 was a blind experiment - due to lack of time we did not try this combination on development and training sets. It is interesting to notice that this RUN achieves the best AUC (50%), while maintaining full recall (like RUN 2). The preliminary conclusion appears to be that system A produces a better ranking, which, when combined with the more complete output of system B, results in a better AUC. While system A has been specifically optimized for the IMT task with task-specific heuristics, system B provides a fairly generic implementation of a naive Bayes multiclass classifier, which therefore does not need a very detailed description. In the rest of this section we provide more information about System A.

As a first approach, we used a pattern matcher giving high scores to every occurrence of an exact match, and lower scores to every occurrence of a word-submatch, using the PSI-MI dictionary of experimental methods [9] as our standard. No 'stop word' list was used, except for removing the prepositions *of* and *in* which occur in many terms and synonyms. The inclusion of submatches led to overgeneration (increased recall but low precision). Using only full matches led to very low recall. As an intermediate level between full match and word-based submatch, we also used a subset approach: if more than three words of a term or a synonym from the PSI-OBO dictionary appear in a ten word observation window, a mid-range score is given for each occurrence. We observed that some submatch words are contained in many different experimental methods (they do not discriminate well) and at the same time many submatch words very often do not indicate a method mention. For example, method 0231 has the term name *mammalian protein interaction trap*, which means that every occurrence of the word *protein* assigns a score to this method.

To respond to these observations, a statistical method can be used. We use, on the one hand conditional probabilities for the method given a word $p(method \mid term\ word)$ and, on the other hand the conditional probability that a given submatch word occurs in a document where the corresponding term identifier has been effectively assigned by the annotator: $p(term\ word = yes \mid word, document)$. We informally refer to the latter probability as termness. We first use the statistical model, $p(method \mid word) * termness(word)$ for all words that are matches or submatches of the terms given in the PSI-MI dictionary, and further for all words, irrespective of whether they appear in the PSI dictionary, whenever $p(method|word)$ and $termness(word)$ are above 10%, and whenever the word is used in at least 5 training documents. We have obtained considerably better results when using the statistical model also on all words, including non-term words. The lists containing words which have a high probabilities to be associated with a given method are not obviously interpretable by the non-expert, although some of the inherent knowledge they contain are clear hints. An excerpt of frequent words indicating experimental methods at high probability is given in table 4.

IAT Task

The ODIN system is being developed within the scope of the OntoGene project, as a collaboration between the OntoGene group at the University of Zurich and the NITAS/TMS group (Text Mining Services) of Novartis Pharma AG. The purpose of the system is to allow a human annotator/curator to leverage upon the result of a text mining system in order to enhance the speed and effectiveness of the annotation process.

The OntoGene system takes as input a document in plain text or a number of supported xml-based formats (including PubMed Central) and processes it with a custom NLP pipeline, which includes Named Entity recognition and relation extraction. Entities which are currently supported include proteins, genes, experimental methods, cell lines, species. Entities detected in the input document are disambiguated with respect to a reference database (UniProt, EntrezGene, NCBI taxonomy, PSI-MI ontology).

The annotated documents are handed back to the ODIN interface (as pure XML documents), which allows multiple display modalities, plus various selection and modification options. The curator/annotator can view the whole document with in-line annotations highlighted, or can browse the extracted entities and be pointed back to the mentions of the entities within the original document. All entity mentions are entirely editable: the curator can easily add or delete any of them, and also change its extent (i.e. add/remove words to its right or left) with a simple click of the mouse. Different entity views are supported, with sorting capabilities according to different criteria (entity type, entity mention, confidence score, etc.). Selective highlighting of text units (e.g. sentences) containing desired entities (terms or gene identifiers) is supported. Rapid disambiguation can be achieved through manual organism selection. Additionally, extensive logging functionalities are provided. The curation interface is mainly developed as a JavaScript-based web application using the extjs framework. This allows rapid prototyping of views (tables, highlighting, creation of hyperlinks). Visualization is very flexible through CSS and DOM manipulation.

Acknowledgements

The OntoGene group is partially supported by the Swiss National Science Foundation (grants 100014 – 118396/1 and 105315 – 130558/1) and by NITAS/TMS, Text Mining Services, Novartis Pharma AG, Basel, Switzerland.

References

1. Rinaldi F, Schneider G, Kaljurand K, Hess M, Romacker M: **An Environment for Relation Mining over Richly Annotated Corpora: the case of GENIA**. *BMC Bioinformatics* 2006, **7**(Suppl 3):S3, [http://www.biomedcentral.com/1471-2105/7/S3/S3].
2. Rinaldi F, Kappeler T, Kaljurand K, Schneider G, Klenner M, Clematide S, Hess M, von Allmen JM, Parisot P, Romacker M, Vachon T: **OntoGene in BioCreative II**. *Genome Biology* 2008, **9**(Suppl 2):S13, [http://genomebiology.com/2008/9/S2/S13].
3. Schneider G, Kaljurand K, Kappeler T, Rinaldi F: **Detecting protein-protein interactions in biomedical texts using a parser and linguistic resources**. In *Proceedings of CICLING 2009* 2009.
4. Rinaldi F, Schneider G, Kaljurand K, Clematide S, Vachon T, Romacker M: **OntoGene in BioCreative II.5**. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2010, **7**(3):472–480.
5. Alias-i: **LingPipe**. [http://alias-i.com/lingpipe].
6. Kappeler T, Kaljurand K, Rinaldi F: **TX Task: Automatic Detection of Focus Organisms in Biomedical Publications**. In *Proceedings of the BioNLP workshop, Boulder, Colorado* 2009.
7. Daumé III H: **Notes on CG and LM-BFGS Optimization of Logistic Regression** 2004. [Paper available at http://pub.hal3.name#daume04cg-bfgs, implementation available at http://hal3.name/megam/].
8. Alex B, Grover C, Haddow B, Kabadjov M, Klein E, Matthews M, Roebuck S, Tobin R, Wang X: **Assisted Curation: Does Text Mining Really Help**. In *BIOCOMPUTING 2008. Proceedings of the Pacific Symposium on Biocomputing*. Edited by Altman RB, Dunker AK, Hunter L, Murray T, Klein TE, Kohala Coast, Hawaii, USA 2008[http://psb.stanford.edu/psb-online/proceedings/psb08/alex.pdf].
9. Hermjakob H, Montecchi-Palazzi L, Bader G, Wojcik J, Salwinski L, Ceol A, Moore S, Orchard S, Sarkans U, von Mering C, Roechert B, Poux S, Jung E, Mersch H, Kersey P, Lappe M, Li Y, Zeng R, Rana D, Nikolski M, Husi H, Brun C, Shanker K, Grant SG C Sander, Bork P, Zhu W, Pandey A, Brazma A, Jacq B, Vidal M, Sherman D, Legrain P, Cesareni G, Xenarios I, Eisenberg D, Steipe B, Hogue C, R A: **The HUPO PSI's molecular interaction format - a community standard for the representation of protein interaction data**. *Nat. Biotechnol* 2004, **22**:177–183.